CLAIMS

What is claimed is:

Sub C3

1. A method for detecting similar documents comprising the steps of:

obtaining a document;

filtering the document to obtain a filtered document;

determining a document identifier for the filtered document and a hash value for the filtered document;

generating a tuple for the filtered document, the tuple comprising the document identifier for the filtered document and the hash value for the filtered document;

comparing the tuple for the filtered document with a document storage structure comprising a plurality of tuples, each tuple in the plurality of tuples representing one of a plurality of documents, each tuple in the plurality of tuples comprising a document identifier and a hash value; and

determining if the tuple for the filtered document is clustered with another tuple in the document storage structure, thereby detecting if the document is similar to another document represented by the another tuple in the document storage structure.

2. A method as in claim 1, wherein the step of filtering comprises parsing the document, and wherein the filtered document comprises a token stream, the token stream comprising a plurality of tokens.

3. A method as in claim 2, wherein the step of filtering further comprises retaining a token in the token stream as a retained token according to at least one token threshold.

4. A method as in claim 3, wherein the step of filtering further comprises arranging the retained tokens in the token stream to obtain an arranged token stream.

5. A method as in claim 3, wherein the step of determining the hash value for the filtered document comprises determining the hash value by processing individually each retained token in the token stream.

6. A method as in claim 2, wherein the step of filtering further comprises:

determining a score for each token in the token stream;

comparing the score for each token to a first token threshold; and

modifying the token stream by removing each token having a score not satisfying the first token threshold and retaining each token as a retained token having a score satisfying the first token threshold.

7. A method as in claim 6, wherein the step of filtering further comprises:

comparing the score for each retained token to a second token threshold; and

modifying the token stream by removing each retained token having a score not satisfying the second token threshold and retaining each retained token having a score satisfying the second token threshold.

8. A method as in claim 2, wherein the step of filtering further comprises removing from the token stream at least one token corresponding to a stop word.

9. A method as in claim 2, wherein the step of filtering further comprises removing a token from the token stream if the token is a duplicate of another token in the token stream.

10. A method as in claim 2, wherein the step of filtering further comprises removing a token from the token stream if the token is either a very frequent token or a very infrequent token.

11. A method as in claim 2, wherein the step of filtering comprises removing at least one token from the token stream.

12. A method as in claim 1, wherein the step of filtering comprises removing formatting from the document.

13. A method as in claim 1, wherein the step of filtering uses collection statistics for filtering the document.

14. A method as in claim 13, wherein the collection statistics pertain to the plurality of documents.

15. A method as in claim 1, wherein the step of determining the hash value for the filtered document comprises using a hash algorithm to determine the hash value, the hash algorithm having an approximately even distribution of hash values.

16. A method as in claim 1, wherein the step of determining the hash value for the filtered document comprises using a standard hash algorithm to determine the hash value.

17. A method as in claim 1, wherein the step of determining the hash value for the filtered document comprises using a secure hash algorithm to determine the hash value.

18. A method as in claim 1, wherein the step of determining the hash value for the filtered document comprises using hash algorithm SHA-1 to determine the hash value.

19. A method as in claim 1, wherein the document storage structure comprises a hash table.

20. A method as in claim 1, wherein the document storage structure comprises a tree.

21. A method as in claim 20, wherein the tree comprises a binary tree.

22. A method as in claim 21, wherein the binary tree comprises a binary balanced tree.

23. A method as in claim 1, wherein the document storage structure comprises a hash table and at least one tree.

24. A method as in claim 1, wherein the step of comparing comprises inserting the tuple into the document storage structure.

25. A method as in claim 1, wherein the document storage structure comprises a hash table, the hash table comprising a plurality of bins, each bin of the hash table comprising at least one tuple of the plurality of tuples, and

wherein the step of determining if the tuple is clustered with another tuple comprises determining if the tuple is co-located with another tuple at a bin of the hash table.

26. A method as in claim 1, wherein the document storage structure comprises a tree, the tree comprising a plurality of branches, each bucket of the tree comprising at least one tuple of the plurality of tuples, and

wherein the step of determining if the tuple is clustered with another tuple comprises determining if the tuple is co-located with another tuple in a bucket of the tree.

27. A computer for performing the method of claim 1.

28. A computer-readable medium having software for performing the method of claim 1.

29. An apparatus for detecting similar documents comprising:

means for obtaining a document;

means for filtering the document to obtain a filtered document;

means for determining a document identifier for the filtered document and a hash value for the filtered document;

means for generating a tuple for the filtered document, the tuple comprising the document identifier for the filtered document and the hash value for the filtered document;

means for comparing the tuple for the filtered document with a document storage structure comprising a plurality of tuples, each tuple in the plurality of tuples representing one of a plurality of documents, each tuple in the plurality of tuples comprising a document identifier and a hash value; and

means for determining if the tuple for the filtered document is clustered with another tuple in the document storage structure, thereby detecting if the document is similar to another document represented by the another tuple in the document storage structure.

30. A method for detecting similar documents comprising the steps of:

obtaining a document;

parsing the document to remove formatting and to obtain a token stream, the token stream comprising a plurality of tokens;

retaining only retained tokens in the token stream by using at least one token threshold;

arranging the retained tokens to obtain an arranged token stream;

processing in turn each retained token in the arranged token stream using a hash algorithm to obtain a hash value for the document;

generating a document identifier for the document;

forming a tuple for the document, the tuple comprising the document identifier for the document and the hash value for the document;

inserting the tuple for the document into a document storage tree, the document storage tree comprising a plurality of tuples, each tuple located at a bucket of the document storage tree, each tuple in the plurality of tuples representing one of a plurality of documents, each tuple in the plurality of tuples comprising a document identifier and a hash value; and

determining if the tuple for the document is co-located with another tuple at a same bucket in the document storage tree, thereby detecting if the document is similar to another document represented by the another tuple in the document storage tree.

31. A computer for performing the method of claim 30.

32. A computer-readable medium having software for performing the method of claim 30.

33. An apparatus for detecting similar documents comprising:

means for obtaining a document;

means for parsing the document to remove formatting and to obtain a token stream, the token stream comprising a plurality of tokens;

means for retaining only retained tokens in the token stream by using at least one token threshold;

means for arranging the retained tokens to obtain an arranged token stream;

means for processing in turn each retained token in the arranged token stream using a hash algorithm to obtain a hash value for the document;

means for generating a document identifier for the document;

means for forming a tuple for the document, the tuple comprising the document identifier for the document and the hash value for the document;

means for inserting the tuple for the document into a document storage tree, the document storage tree comprising a plurality of tuples, each tuple located at a bucket of the document storage tree, each tuple in the plurality of tuples representing one of a plurality of documents, each tuple in the plurality of tuples comprising a document identifier and a hash value; and

means for determining if the tuple for the document is co-located with another tuple at a same bucket in the document storage tree, thereby detecting if the document is similar to another document represented by the another tuple in the document storage tree.

34. A method for detecting similar documents comprising the steps of:

determining a hash value for a document;

accessing a document storage structure comprising a plurality of hash values, each hash value in the plurality of hash values representing one of a plurality of documents; and

determining if the hash value for the document is equivalent to another hash value in the document storage structure, thereby detecting if the document is similar to another document represented by the another hash value in the document storage structure.

35. A computer for performing the method of claim 34.

36. A computer-readable medium having software for performing the method of claim 34.

37. An apparatus for detecting similar documents comprising:

means for determining a hash value for a document;

means for accessing a document storage structure comprising a plurality of hash values, each hash value in the plurality of hash values representing one of a plurality of documents; and

means for determining if the hash value for the document is equivalent to another hash value in the document storage structure, thereby detecting if the document is similar to another document represented by the another hash value in the document storage structure.

38. A method for detecting similar documents comprising the step of:

comparing a document to a plurality of documents in a document collection using a hash algorithm and collection statistics to detect if the document is similar to any of the documents in the document collection.

39. A method as in claim 38, wherein the collection statistics pertain to the document collection.

40. A computer for performing the method of claim 38.

41. A computer-readable medium having software for performing the method of claim 38.

42. An apparatus for detecting similar documents comprising:

means for comparing a document to a plurality of documents in a document collection using a hash algorithm and collection statistics to detect if the document is similar to any of the documents in the document collection.